Stockholms universitet
Statistiska institutionen
Raul Cano

SKRIVNINGSDATUM: 09-01-2018

Skriftlig tentamen i **Undersökningsmetodik** (4,5 hp), ingående som moment 1 i kursen **Regressionsanalys och undersökningsmetodik, 15 hp.**

Skrivtid: 5 timmar
Hjälpmedel: Miniräknare utan lagrade formler eller lagrad text. Vidhäftade formel- och tabellblad (obs! vidhäftas endast de tabellsidor som behövs för den här tentamen).
Tentamensgenomgång och återlämning: måndagen den 22 januari, kl. 16.00 i B705.
Därefter kan skrivningarna hämtas på studentexpeditionen, plan 7 i B-huset.

Tentamen består av fem uppgifter som kan ge totalt 100 poäng. För betyget A gäller 90-100 p., för betyget B gäller 80-89 p., för betyget C gäller 70-79 p., för betyget D gäller 60-69 p., för betyget E gäller 50-59 p., för betyget Fx gäller 40-49 p. och för betyget F gäller 0-39 p. För detaljerade betygskriterier se kursbeskrivningen på kurshemsidan.
För full poäng på en uppgift krävs fullständiga och väl motiverade lösningar.

Uppgift 1: (20 poäng)
En population innehåller $N = 5$ element

$x_1 = 6, \; x_2 = 10, \; x_3 = 4, \; x_4 = 2, \; x_5 = 10$

Ur denna population skall man göra ett obundet slumpmässigt urval (OSU) om $n = 2$ element.

a). Beräkna samplingfördelningen för den stokastiska variabeln $\overline{X}$. (10 poäng)

b). Beräkna $E(\overline{X})$ och $V(\overline{X})$. (10 poäng)

Uppgift 2: (30 poäng)
Uppgiften handlar om den lilla republiken Slovenien. För att bestämma en skattning av befolkningen 31 dec 2017 gjordes ett obundet slumpmässigt urval utan återläggning av 5 kommuner bland alla Sloveniens 50 kommuner. Man tog reda på den definitiva befolkningen vid denna tidpunkt i dessa kommuner och man hade total tillgång på motsvarande uppgifter på befolkningen år 2002 (den lilla republiken hade genomfört en folkräkning år 2002). Total i den lilla republiken uppgick befolkningen 31 dec 2002 till 1 964 tusentals personer (alltså 1 964 000 personer). Följande resultat erhölls, alla siffror i tusentals personer:

Kommun	31 dec 2002	31 dec 2017
Gorizia E	20	18
Sava C	58	60
Karst N	30	30
Mura W	10	12
Savinja S	20	22

a). Beräkna kvotskattningen av den totala befolkningen i Slovenien 31 dec 2017 med befolkning 31 dec 2002 som hjälpinformation. (5 poäng)

b). Beräkna den skattade variansen för kvotskattningen av den totala befolkningen i Slovenien 31 dec 2017 med befolkning 31 dec 2002 som hjälpinformation. (10 poäng)

c). Beräkna ett 95%-igt konfidensintervall för den totala befolkningen i Slovenien 31 dec 2017 med hjälp av kvotskattningen. (5 poäng)

d) Beräkna ett 95%-igt konfidensintervall för den totala befolkningen i Slovenien 31 dec 2017 med hjälp av den vanliga OSU-skattningen. (10 poäng)

Uppgift 3: (10 poäng)

När är kvotskattningen bättre än den vanliga OSU-skattningen? (10 poäng)

Uppgift 4: (30 poäng)
I den lilla republiken Slovenien vill man göra en undersökning för att ta reda på vilka utgifter för boende som studenterna vid Ljubljana universitet har i genomsnitt per månad (utgifter i svenska kronor). Populationen har stratifierats i tre stratum efter akademitillhörighet. En stickprovsundersökning genomfördes för n=300 individer. Följande resultat erhölls då:

Stratum	N_i	n_i	$\bar{x}_i$	s_i
Handelshögskolan	1200	200	2825	300
Fakultet för humaniora, utbildning och samhällsvetenskap	1000	50	2400	290
Fakultet för Juridik, psykologi och socialt arbete	900	50	2700	320

a). Beräkna ett 95%-igt konfidensintervall för den genomsnittliga boendekostnaden per månad. (10 poäng)

En statistiker tog del av undersökningen och undrade hur stickprovet hade allokerats. Det visade sig att det fanns finansiella medel för att göra om undersökningen.

b). Från sin erfarenhet på detta område säger statistikern att standardavvikelserna i populationen borde vara lika med $\sigma_1 = 350$, $\sigma_2 = 200$ och $\sigma_3 = 300$, där σ_i står för standardavvikelsen i stratum nummer i. Beräkna hur de trehundra observationerna ska allokeras i detta fall. (10 poäng)

En annan statistiker säger att det är rimligare att förutsätta att standardavvikelserna är lika.

c). Beräkna hur de trehundra observationerna ska allokeras i detta fall. (10 poäng)

Uppgift 5: (10 poäng)
Antag att vi vill skatta P = andelen som skulle rösta med regeringsalliansen om det vore val idag. Hur stort urval måste vi dra om vi skall göra ett OSU utan återläggning bland "Svenska folket 16 år och äldre" samt vill ha en felmarginal på högst 3 % och en konfidensgrad på 95 %? (10 poäng)

Formelsamling undersökningsmetodik

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \hat{\tau} = N\bar{X}$$

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}}{n-1} = \frac{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}{n(n-1)}$$

Beräkning av stickprovsstorlek:

$$n \geq \frac{N\sigma^2}{D^2(N-1) + \sigma^2}$$

Stratifierat urval:

$$\bar{X}_{st} = \sum_{i=1}^{L} W_i \bar{X}_i \qquad V(\bar{X}_{st}) = \sum_{i=1}^{L} W_i^2 V(\bar{X}_i) \qquad \text{där} \quad W_i = \frac{N_i}{N}$$

Optimal allokering:

$$n_i = n\frac{N_i \sigma_i}{\sum_{j=1}^{L} N_j \sigma_j}$$

Skattning av medelvärde samt proportion per element:

$$\bar{X}_{kvot} = \frac{\sum_{i=1}^{n} \tau_i}{\sum_{i=1}^{n} m_i} \qquad \bar{X}_{VVR} = N\frac{\bar{\tau}}{M} \qquad p_{kvot} = \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} m_i} \qquad P_{VVR} = N\frac{\bar{a}}{M}$$

Punktskattning OSU	*Varians m.å.*	*Variansskattning m.å.*	*Varians u.å.*	*Variansskattning u.å.*
$\bar{X}$	$\frac{\sigma^2}{n}$	$\frac{s^2}{n}$	$\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$	$\frac{s^2}{n}\left(1-\frac{n}{N}\right)$
$\hat{\tau}$	$N^2 \cdot \frac{\sigma^2}{n}$	$N^2 \cdot \frac{s^2}{n}$	$N^2 \cdot \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$	$N^2 \cdot \frac{s^2}{n}\left(1-\frac{n}{N}\right)$
p	$\frac{P(1-P)}{n}$	$\frac{p(1-p)}{n-1}$	$\frac{P(1-P)}{n}\left(\frac{N-n}{N-1}\right)$	$\frac{p(1-p)}{n-1}\left(1-\frac{n}{N}\right)$
$\hat{A}$	$N^2 \cdot \frac{P(1-P)}{n}$	$N^2 \cdot \frac{p(1-p)}{n-1}$	$N^2 \cdot \frac{P(1-P)}{n}\left(\frac{N-n}{N-1}\right)$	$N^2 \cdot \frac{p(1-p)}{n-1}\left(1-\frac{n}{N}\right)$

Tillägg till formelsamling undersökningsmetodik

Skattning av τ_X. Urval OSU

$$\hat{\tau}_{kvot} = \hat{R} \cdot \tau_Z = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} z_i} \cdot \tau_Z$$

$$\hat{V}\left(\hat{\tau}_{kvot}\right) = N^2 \left(\frac{N-n}{nN}\right) \frac{\sum_{i=1}^{n} \left(x_i - \hat{R} z_i\right)^2}{n-1}$$

$$\sum_{i=1}^{n} \left(x_i - \hat{R} z_i\right)^2 = \sum_{i=1}^{n} x_i^2 + \hat{R}^2 \sum_{i=1}^{n} z_i^2 - 2\hat{R} \sum_{i=1}^{n} x_i z_i$$

Skattning av μ_X. Urval OSU.

$$\hat{\mu}_{reg} = \overline{x} + b\left(\mu_Z - \overline{z}\right)$$

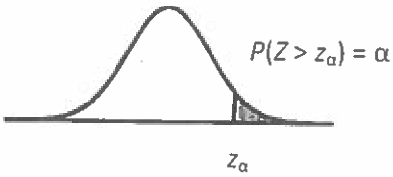
$$b = \frac{\sum_{i=1}^{n} \left(z_i - \overline{z}\right)\left(x_i - \overline{x}\right)}{\sum_{i=1}^{n} \left(z_i - \overline{z}\right)^2}$$

$$\hat{V}\left(\hat{\mu}_{reg}\right) = \left(\frac{N-n}{nN}\right)\left(\frac{1}{n-2}\right)\left[\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2 - b^2 \sum_{i=1}^{n} \left(z_i - \overline{z}\right)^2\right]$$

TABELL 2. Normalfördelningens kvantiler, standardiserad

$Z \in N(0, 1)$. Vilket värde har z_α om $P(Z > z_\alpha) = \alpha$ där α är en given sannolikhet.

Utnyttja även $\Phi(-z) = 1 - \Phi(z)$ för $P(Z \leq -z_\alpha)$.



α	z_α
0,25	0,6745
0,10	1,2816
0,05	1,6449
0,025	1,9600
0,010	2,3263
0,005	2,5758
0,0025	2,8070
0,0010	3,0902
0,0005	3,2905
0,00025	3,4808
0,00010	3,7190
0,00005	3,8906
0,000025	4,0556
0,000010	4,2649
0,000005	4,4172

2

Statistiska institutionen

Stockholms universitet

Rättningsblad

Datum: 09/01/18

Sal: Uggleviksssalen

Tenta: Undersökningsmetodik

Kurs: Regressionsanalys och undersökningsmetodik

ANONYMKOD: UND-JKN-TBW

☒ Jag godkänner att min tenta får läggas ut anonymt på hemsidan som studentsvar.

OBS! SKRIV ÄVEN PÅ BAKSIDAN AV SKRIVBLADEN

Markera besvarade uppgifter med kryss

1	2	3	4	5	6	7	8	9	Antal inl. blad
X	X	X	X	X					4 3
Lär.ant. 20p	[illegible] 20p	10p	30p	10p					

POÄNG	BETYG	Lärarens sign.
90p	A	RC

SU, STATISTIK

Skrivsal: Ugglevik **Anonymkod:** UND-JKN-TBW **Blad nr:** 1

1) 20p

UPPGIFT 1

$N = 5 \qquad n = 2$

I POPULATIONEN

$$\mu = \frac{(6+10+4+2+10)}{5} = 6{,}4$$

$$\sigma^2 = (6-6{,}4)^2 + (10-6{,}4)^2 + (4-6{,}4)^2 + (2-6{,}4)^2 + (10-6{,}4)^2 \Rightarrow$$

$$\frac{0{,}16 + 12{,}96 + 5{,}76 + 19{,}36 + 12{,}96}{5} = 10{,}24$$

a) Hur många urval kan vi dra?

$$\binom{N}{n} = \frac{5!}{2!\,3!} = \frac{5 \cdot 4}{2} = \underline{10}$$

	Urval	$\bar{x}$	s^2
1	6,10	8	8
2	6,4	5	2
3	6,2	4	8
4	6,10	8	8
5	2,4	3	2
6	2,10	6	32
7	2,10	6	32
8	4,10	7	18
9	4,10	7	18
10	10,10	10	0
Σ		64	128
Medelvärde		6,4	12,8

$$\bar{x} = \frac{6+10}{2} = 8$$

$$\bar{x} = \frac{6+4}{2} = 5$$

Hur jag räknat värdet $\bar{x}$

$$s^2 = \frac{(6-10)^2}{2} = 8$$

$$s^2 = \frac{(6-4)^2}{2} = 2$$

Hur jag räkna värdet s^2

$\Rightarrow$

Svar på a)

Sampling fördelning för $\bar{x}$

$\bar{x}$	8	5	4	3	6	7	10	
$P(\bar{x})$	0,2	0,1	0,1	0,1	0,2	0,2	0,1	1

R

b) $E(\bar{x}) = 8\cdot 0{,}2 + 5\cdot 0{,}1 + 4\cdot 0{,}1 + 3\cdot 0{,}1 + 6\cdot 0{,}2 + 7\cdot 0{,}2 + 10\cdot 0{,}1 = 1{,}6 + 0{,}5 + 0{,}4 + 0{,}3 + 1{,}2 + 1{,}4 + 1$

$\Rightarrow \boxed{6{,}4}$ R

$V(\bar{x}) = (8-6{,}4)^2\cdot 0{,}2 + (5-6{,}4)^2\cdot 0{,}1 + (4-6{,}4)^2\cdot 0{,}1 + (3-6{,}4)^2\cdot 0{,}1 + (6-6{,}4)^2\cdot 0{,}2 + (7-6{,}4)^2\cdot 0{,}2 + (10-6{,}4)^2\cdot 0{,}1 = 0{,}512 + 0{,}196 + 0{,}576 + 1{,}156 + 0{,}032 + 0{,}072 + 1{,}296 \Rightarrow \boxed{3{,}84}$ R

Om vi sätter in $\frac{10{,}24}{2}\cdot\left(\frac{5-2}{5-1}\right) = 5{,}12\cdot 0{,}75 = 3{,}84$

Och vi kan se från itargare att våra $\bar{x}$ värden i genomsnitt är lika med μ populationens medelvärde vilket innebär att våra skattningar inte innehåller några systematiska fel dvs de är väntevärdesriktiga.

R

Svar $E(\bar{x}) = 6{,}4$ $V(\bar{x}) = 3{,}84$

SU, STATISTIK

Skrivsal: Ugglevik Anonymkod: UND-7KV-TBW Blad nr: 2

2) 20 p

UPPGIFT 2

OSU - U.Å $n=5$ $N=50$ Totalbefolkning 2002 1964 tusen

a) Hjälpvariabel $\sum_{i=1}^{n} z_i = 138$ (i tusental)

Undersökningsvariabel $\sum_{i=1}^{n} x_i = 142$ (i tusental)

Kvotskattning $= \hat{R} \cdot T_z = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} z_i} \cdot T_z$

$\hat{R} = \frac{142}{138} = 1{,}028985507$

$T_z = 1964$ (i tusental)

$\boxed{\hat{T}_{kvot} = \frac{142}{138} \cdot 1964 = 2020{,}927536 \text{ (i tusental)}}$ R

Svar: Kvotskattningen av den totala befolkningen år 2017 är ungefär 2020,93 (i tusen)

b) $\hat{V}(\hat{T}_{kvot}) = N^2\left(\frac{N-n}{nN}\right) \sum_{i=1}^{n} \frac{(x_i - \hat{R}z_i)^2}{n-1}$

Vi behöver ställa upp följande tabell:

	x_i^2	z_i^2	$x_i z_i$
	324	400	360
	3600	3364	3480
	900	900	900
	144	100	120
	484	400	440
Σ	5452	5164	5300

$V(\hat{T}_{kvot})$ $50^2\left(\frac{50-5}{50 \cdot 5}\right) \cdot \frac{12{,}45362}{5-1} =$

$2500 \cdot 0{,}18 \cdot 3{,}113405 \Rightarrow$

$1401{,}03225 \approx 1401$

$\sum (x_i - \hat{R}z_i)^2 \Rightarrow$

$5452 + \left(\frac{142}{138}\right)^2 \cdot 5164 - 2 \cdot \left(\frac{142}{138}\right) \cdot 5300$

$\Rightarrow 5452 + 5467{,}7 - 10907{,}24638 \Rightarrow \underline{\underline{12{,}45362}}$

Svar: Den skattade variansen blir 1401.

R

c) Konfidensintervall:

Vårat n är litet så vi kan inte förlita oss på CGS, vi får anta normalfördelning.

Felmarginalen: $1,96 \cdot \sqrt{V}$ (root) { 1,96 pga $Z_{0,025}$ från tabell }

$1,96 \cdot \sqrt{1401} = 1,96 \cdot 37,42993454 = 73,36267171$

$2020,93 \pm 73,36267171$

$[1947,667328 ; 2094,292672]$

Svar: Med 95% konfidens ligger det sanna värdet i intervallet ovan.

R

d) 95%-igt konfidensintervall:

$\bar{x} = 28,4$

$$s^2 = \frac{\sum x_i^2 - n \cdot \bar{x}^2}{n-1} = \frac{5452 - 5 \cdot (28,4)^2}{4} = \frac{5452 - 4032,8}{4}$$

$= 354,8$

$$\frac{354,8}{5}\left(1 - \frac{5}{50}\right) = 70,96 \cdot 0,9 \Rightarrow 63,864$$

$28,4 \pm 1,96\sqrt{63,864} \Rightarrow 28,4 \pm 15,6633114$

$[12,7366886 ; 44,0633114]$

F

Svar: Med 95% konfidens ligger det sanna värdet i intervallet ovan.

F

F

FEL METOD (SE FACIT)

SU, STATISTIK

Skrivsal: Ugglevik Anonymkod: UND-JKN-TBW Blad nr: 3

3) 10p

UPPGIFT 3

Kvotskattningen är bättre när vi har hjälpinformation tillgänglig. Hjälpinformationen dvs variabeln z ska vara korrelerad med vår undersökningsvariabel x. Om x och z är en rät linje genom origo så passar kvotskattningen bäst. Helst ska de vara starkt och positivt korrelerade för att skattningen ska bli bättre. Men om vi endast har en rät linje som inte går igenom origo passar regressionskattningen bättre.

R

R

BRA!

Sammanfattningsvis är en kvotskattning att föredra när vi har en rät linje genom origo dvs de är korrelerade (starkt, positivt) över en vanlig OSU.

Korrelation kan vara mellan -1 och 1 och mäter linjära samband.

Exempelvis: Om vi vill skatta en persons inkomst så kan utbildning vara en bra hjälpinformation, och har vi den blir troligtvis våra skattningar bättre än vanlig OSU utan hjälpinformation.

VÄND BLAD

⇒

UPPGIFT 4 4) 30p

$n = 300$

N_i	n_i	$\bar{x}_i$	s_i	w_i
1200	200	[illegible]	300	0,3870967742
1000	50	2400	290	0,3225806452
900	50	2700	320	0,2903225806
Σ 3100				

$\Rightarrow \frac{N_i}{N} = w_i$

a) $\bar{x}_{st} = w_i \bar{x}_i = 0,387\ldots \cdot 2825 + 0,322\ldots \cdot 2400 + 0,290\ldots \cdot 2700 = 1093,548387 + 774,1935484 + 783,8709677 = \boxed{2651,612903}$ R

$V(\bar{x}_{st}) = w_i^2 \cdot \frac{s^2}{n}\left(1 - \frac{n}{N}\right) = \left(\frac{1200}{3100}\right)^2 \cdot \frac{300^2}{200} \cdot \left(1 - \frac{200}{1200}\right) + \left(\frac{1000}{3100}\right)^2 \cdot \frac{290^2}{50}\left(1 - \frac{50}{1000}\right) + \left(\frac{900}{3100}\right)^2 \cdot \frac{320^2}{50} \cdot \left(1 - \frac{50}{900}\right)$

$\Rightarrow 56,19146722 + 166,274738 + 163,0301764$

$\boxed{\Rightarrow 385,4963579}$ R

95%-igt CI

Från tabell

$\bar{x} \pm z_{\alpha/2} \sqrt{V(\bar{x})} = 2651,612903 \pm 1,96 \cdot \sqrt{385,4963579}$

$\boxed{\begin{array}{l} 2651,612903 \pm 38,48275989 \\ [2613,130143 ; 2690,095663] \end{array}}$ R

Svar: Med 95% konfidens ligger det sanna värdet i intervallet ovan. R

SU, STATISTIK

Skrivsal: Ugglevik Anonymkod: UND-JKN-TRW Blad nr: 4

UPPGIFT 4 FORTSÄTTNING

b) Optimal allokering $\Rightarrow \frac{N_i \sigma_i}{\sum N_j \cdot \sigma_j} \cdot n$

$1200 \cdot 350 = 42\,00\,00$

$1000 \cdot 200 = 20\,00\,00$

$900 \cdot 300 = 27\,0000$

$\Sigma\ 89\,00\,00$

~~$\frac{42\,00\,00}{89\,00\,00} \cdot 300 = 141{,}57 \approx 142$~~

~~$\frac{20\,0000}{890000} \cdot 300 = 67{,}4157 \approx 68$~~

~~$\frac{27\,00\,00}{89\,00\,00} \cdot 300 = 91$~~

~~$142 + 68 + 91 = 300$~~

R

c) Proportionell allokering

$n_i = n \cdot \frac{N_i}{N}$

$300 \cdot \frac{1200}{3100} = 116{,}12 \approx 116$

$300 \cdot \frac{1000}{3100} = 96{,}77 \approx 97$

$300 \cdot \frac{900}{3100} = 87{,}09 \approx 87$

$116 + 97 + 87 = 300$

R

Urvalet dras proportionellt från populationen

$\Rightarrow$ vänd blad

Man kan dra urvalet på tre olika sätt

	Ekonomer	Lärare	Jurister	
Pop	0,3	0,6	0,2	Proportionellt
Urval	0,3	0,6	0,2	

Eller så kan man dra lika ur alla urval

	Ekonomer	Lärare	Jurister	
Pop	500	1000	100	n=150
Urval	50	50	50	lika i vare n:

Eller så kan man köra optimal allokering $\sigma = \sqrt{p(1-p)}$ och då behöver man standardavvikelsen för varje stratum.

UPPGIFT 5

$1{,}96 \cdot \sqrt{\frac{p(1-p)}{n}} \leq 0{,}03$

5) 10p

1. Vi använder oss av största möjliga varians
   $0{,}5 \cdot 0{,}5 = 0{,}25$
   $1{,}96 \cdot \sqrt{\frac{0{,}25}{n}} < 0{,}03$
2. $1{,}96^2 \cdot \frac{0{,}25}{n} < 0{,}03^2$
3. $1{,}96^2 \cdot 0{,}25 < 0{,}03^2 \cdot n$
4. $\frac{1{,}96^2 \cdot 0{,}25}{0{,}03^2} < n$
5. $\frac{0{,}9604}{0{,}0009} = n$

n = 1067,1111 Man avrundar alltid uppåt så svar n = 1068

R

R